

The Intrinsic Fraction of Broad Absorption Line Quasars

Christian Knigge¹*, Simone Scaringi¹, Michael R. Goad² and Christopher E. Cottis²

¹*Department of Physics and Astronomy, University of Southampton, Highfield, SO17 1BJ, UK*

²*Department of Physics and Astronomy, University of Leicester, University Road, LE1 7RH, UK*

ABSTRACT

We carefully reconsider the problem of classifying broad absorption line quasars (BALQSOs) and derive a new, unbiased estimate of the intrinsic BALQSO fraction from the SDSS DR3 QSO catalogue. We first show that the distribution of objects selected by the so-called “absorption index” (AI) is clearly bimodal in log AI, with only one mode corresponding to definite BALQSOs. The surprisingly high BALQSO fractions that have recently been inferred from AI-based samples are therefore likely to be overestimated. We then present two new approaches to the classification problem that are designed to be more robust than the AI, but also more complete than the traditional “balmicity index” (BI). Both approaches yield *observed* BALQSO fractions around 13.5%, while a conservative third approach suggests an upper limit of 18.3%. Finally, we discuss the selection biases that affect our observed BALQSO fraction. After correcting for these biases, we arrive at our final estimate of the *intrinsic* BALQSO fraction. This is $f_{\text{BALQSO}} = 0.17 \pm 0.01$ (stat) ± 0.03 (sys), with an upper limit of $f_{\text{BALQSO}} \simeq 0.23$. We conclude by pointing out that the bimodality of the log AI distribution may be evidence that the BAL-forming region has clearly delineated physical boundaries.

Key words: quasars: absorption lines; methods, statistical, catalogues, surveys

1 INTRODUCTION

Broad absorption line quasars (BALQSOs) are a sub-class of active galactic nuclei (AGN) that exhibit strong, broad and blue-shifted spectroscopic absorption features (Foltz et al. 1990; Weymann et al. 1991; Reichard et al. 2003). Most BALQSOs – the so-called HiBALs – only display absorption troughs in certain high-ionisation lines (e.g. NV $\lambda 1240\text{\AA}$, CIV $\lambda 1549\text{\AA}$, SiIV $\lambda 1397\text{\AA}$), but some – the so-called LoBALs – also show absorption in some low-ionisation lines (most notably MgII $\lambda 2800\text{\AA}$). BALQSOs are predominantly radio-quiet (Stocke et al. 1992; Becker et al. 2001; Shankar, Dai, & Sivakoff 2008), and there are also subtle differences between their continuum and emission line properties and those of “normal” (non-BAL) QSOs (Reichard et al. 2003). However, despite these differences, BALQSOs and non-BAL QSOs appear to be drawn from the same parent population (Reichard et al. 2003).

The simplest and most promising interpretation of the QSO/BALQSO dichotomy is in terms of an orientation effect. This fits in well with unified models, in which orientation is the major factor determining the observational appearance of AGN (e.g. Elvis 2000). It also makes sense physically, since the absorption troughs in BALQSOs have long

been recognised as signatures of fast, large-scale outflows from the central engines. More specifically, blue-shifted absorption is produced when the central continuum and/or emission line source is viewed through outflowing material that scatters photons out of the observer’s line of sight. If the outflow subtends a solid angle $0 < \Omega < 2\pi$, then both BALQSOs and non-BAL QSOs can be accounted for in this picture.

The powerful outflows we observe in BALQSO are an important example of AGN feedback. Such feedback is the key ingredient in theoretical attempts to understand galaxy “downsizing” and may also be responsible for regulating the growth of supermassive black holes, quenching star formation and setting up the $M_{\text{BH}} - \sigma$ and $M_{\text{BH}} - M_{\text{bulge}}$ relations (e.g. Silk & Rees 1998; King 2003; di Matteo, Springel & Hernquist 2005; Scannapieco, Silk & Bouwens 2005). However, despite their fundamental importance, the geometry, kinematics and energetics of BALQSO outflows have remained highly uncertain.

Perhaps the single most important quantity that can be determined empirically regarding BALQSOs is their incidence within the overall QSO population. More specifically, the BALQSO fraction (f_{BALQSO}) is defined as the fraction of QSOs that display BALQSO absorption features. Its significance derives mainly from the fact that it allows a simple,

* E-mail: christian@astro.soton.ac.uk

geometric interpretation: in the context of unified schemes, f_{BALQSO} is the covering fraction of BALQSO outflows.

Until recently, searches for BALQSOs in quasar surveys consistently reported observed BALQSO fractions around 10%–15% (Weymann et al. 1991; Tolea, Krolik, & Tsvetanov 2002; Hewett & Foltz 2003; Reichard et al. 2003). It therefore came as something of a surprise when Trump et al. (2006) reported a significantly higher BALQSO fraction of 26% from the spectroscopic QSO catalogue associated with the 3rd Data Release (DR3) of the Sloan Digital Sky Survey (SDSS; Schneider et al. 2005).

There can be little question that the QSO sample on which the Trump et al. study is based is superior to earlier QSO surveys. However, this is not the reason for their unusually high estimate of f_{BALQSO} . Instead, Trump et al. argue that the “classic” definition of BALQSOs, based on the so-called “balmicity index” (hereafter, BI) is not appropriate for BALQSO classification purposes. Instead, they prefer a different statistic, the so-called “absorption index” (hereafter, AI). The AI is designed to be less strict than the BI, with the result that a significantly higher fraction of QSOs are classified as broad absorption line (BAL) objects. In essence, Trump et al. (2006) argue that BALs can be both weaker and much narrower than has previously been supposed. QSOs containing such features would naturally be excluded from any census involving the classic BI definition.

If Trump et al. are correct, the covering fraction of BALQSO outflows must be much larger than has previously been assumed. Indeed, it has been suggested that their *observed* BALQSO fraction of 26% implies an *intrinsic* BALQSO fraction of $43\% \pm 2\%$ once selection effects are taken into account (Dai, Shankar & Sivakoff 2008). This is about twice the best previous estimates ($22\% \pm 4\%$ [Hewett & Foltz 2003]; $15.9\% \pm 1.4\%$ [Reichard et al. 2003]).

The main goals of the present paper are to take a fresh look at the metrics used for classifying BALQSOs and to derive a new, robust estimate of f_{BALQSO} . It is worth emphasizing from the outset that what we wish to accomplish is to identify a distinct sub-population of QSO of which classic BALQSOs (with $\text{BI} > 0 \text{ km s}^{-1}$) are just the most obvious representatives. This is an important point, because – as effectively argued by Trump et al. (2006) – these classic BALQSOs may just be the tip of the iceberg. Thus the very term “broad absorption line quasar” could be a mis-nomer, since it is possible that the majority of objects belonging to this population could in principle exhibit only weak/narrow absorption features (or even no absorption at all). It could even turn out that a distinct BALQSOs sub-population does not exist: QSOs could simply exhibit a perfectly continuous and smooth distribution of absorption characteristics, with classic BALQSOs occupying the arbitrarily defined extreme tail of this distribution. As we shall see, there is, in fact, evidence that BALQSOs do form a distinct sub-population. With this in mind, we will use the term BALQSO throughout this paper to denote members of this sub-population, regardless of whether they are identified as such by any given metric. The goal, in fact, is to find ways of quantifying the size of this population in a way that is simultaneously robust (i.e. does not produce many false positives) and complete (i.e. does not miss many true members).

In Section 2, we introduce and compare the widely used

AI and BI metrics for identifying BALQSOs. In Section 3, we show that there is clear evidence for bimodality in the log AI distributions of Trump et al.’s BALQSO candidates, with “classic” BALQSOs (with positive BI) preferentially occupying one mode of the distribution. In Section 4, we present several concrete examples of problematic classifications obtained with *both* standard metrics. In Section 5, we present two new approaches to the classification problem, which are designed to be more robust than the AI, but more complete than the BI. In Section 6, we correct the *observed* BALQSO fractions produced by our new approaches for selection effects and obtain our final estimate of the *intrinsic* BALQSO fraction. Finally, in Section 7, we discuss our results and present our conclusions.

2 HOW BROAD IS BROAD? METRICS FOR IDENTIFYING BROAD ABSORPTION LINE QUASARS

The BI (Weymann et al. 1991) was the first quantitative metric used to identify BALQSOs within QSO surveys. Until the introduction of the AI (see below), the BI remained the standard way to classify objects as BALQSOs. Given a continuum-normalised spectrum in the vicinity of a spectral line, the BI is defined numerically as

$$\text{BI} = - \int_{25000}^{3000} \left[1 - \frac{f(v)}{0.9} \right] C dv. \quad (1)$$

Here, the limits of the integral are in units of km s^{-1} , and $f(v)$ is the normalised flux as a function of velocity displacement from line centre.¹ The constant $C = 0$ everywhere, unless the normalised flux has satisfied $f_c(v) < 0.9$ continuously for at least 2000 km s^{-1} ; at this point it is switched to $C = 1$ until $f(v) > 0.9$ again. Based on this definition, objects are classified as BALQSOs if their $\text{BI} > 0 \text{ km s}^{-1}$.

Physically, the idea behind the BI is to count as BALs only absorption troughs that are definitely real (hence the requirement that $f(v) < 0.9$), definitely broad (hence the demand that troughs must be broader than 2000 km s^{-1} in order to count) and significantly blue-shifted (hence the lower limit of 3000 km s^{-1} on the integral). The main attraction of the BI as a classification tool is that it tends to produce very “clean” BALQSO samples. Indeed, it is hard to imagine a non-BAL QSO being assigned a positive BI unless its spectrum is either very noisy, suffers from a misplaced continuum, or has been assigned an erroneous redshift. However, the conservative nature of the BI also means that BALQSO samples based on it may be seriously incomplete. There is certainly no compelling reason to think that somewhat weaker, narrower and/or less-blue-shifted BALs than recognised by the BI should not exist.

This issue was already recognised by Weymann et al. (1991) and provided the motivation for the introduction of the AI, initially by Hall et al. (2002, here purely as a means

¹ It is worth noting that in the original definition of the BI by Weymann et al. (1991), $f(v)$ is normalised relative to the underlying continuum, whereas other authors, including Trump et al. (2006), normalize relative to a best-fitting continuum plus emission line template.

of identifying systems showing evidence of absorption). The definition of the AI ultimately adopted by Trump et al. (2006) is

$$\text{AI} = \int_0^{29000} [1 - f(v)] C' dv, \quad (2)$$

where $f(v)$ is the normalized flux obtained after dividing the data by the best-fitting emission-line-plus-continuum QSO template. The constant $C' = 1$ in all regions where $f(v) < 0.9$ continuously for at least 1000 km s^{-1} and $C' = 0$ otherwise. Also, only regions containing at least one data point significantly below the underlying continuum are included in the calculation. This ensures that only true absorption features are assigned positive AI. The two key differences that allow some objects with $\text{BI} = 0 \text{ km s}^{-1}$ to achieve $\text{AI} > 0 \text{ km s}^{-1}$ are that (i) the AI includes regions within 3000 km s^{-1} of line centre (and also regions beyond $25,000 \text{ km s}^{-1}$), and (ii) the AI includes objects with much narrower absorption troughs than the BI. The remaining difference is associated with the absence of the factor 0.9 in Equation 2 (compared to Equation 1). This change was made to the definition of the AI in order to allow a clear interpretation: the AI is the combined equivalent width of all absorption troughs in a given line that are located bluewards of line centre, deeper than 0.9 of the continuum, and at least 1000 km s^{-1} wide.

Note that both the AI and the BI can be sensitive to the type of spectrum from which they are measured. For example, an apparently broad absorption trough in a low-resolution spectrum may break up into multiple narrow troughs when observed at higher resolution. Conversely, noise spikes may artificially break up a single trough, so that a true BAL could be assigned zero AI/BI in a noisy spectrum. Throughout this paper, we will use Trump et al.’s (2006) AI/BI estimates for objects in the SDSS DR3 QSO catalogue. The health warning “as derived from its SDSS spectrum” should thus implicitly be added to the AI/BI estimates we use for each QSO.

It is obvious that if BALQSOs are classified on the basis of the less restrictive AI, the resulting BALQSO fraction will be higher than if the BI were used. However, it is not obvious *a priori* that objects selected solely on the basis of having $\text{AI} > 0 \text{ km s}^{-1}$ (i.e. including those with $\text{BI} = 0 \text{ km s}^{-1}$) constitute a single population. The problem is that a wide variety of non-BAL absorption features are commonly seen in QSOs and other AGN. These typically narrower features can be due to absorption at an intermediate redshift along the line of sight to the QSO, absorption within the host galaxy, or intrinsic absorption close to the QSO (including the so-called mini-BALS and associated absorption features) whose origin remains poorly understood and could conceivably be linked to the broad absorption lines. It is therefore extremely difficult to say if any particular QSO containing an “intermediate” width absorption trough ($1000 \text{ km s}^{-1} \lesssim \Delta v \lesssim 3000 \text{ km s}^{-1}$) should be classified as a BALQSO or not. Roughly speaking, the BI metric does not consider *any* such objects to be genuine BALQSOs, whereas the AI metric labels *all* such objects as BALQSOs. In the following section, we will present statistical evidence that the AI metric, in particular, is far too permissive in this respect.

3 THE BIMODAL LOG(AI) DISTRIBUTION OF AI-SELECTED QSOs

Using the definitions above, Trump et al. (2006) calculated AIs and BIs for all 11,611 QSOs in the SDSS DR3 sample. In Figure 1, we show as a black histogram the log AI distribution of the 3182 QSOs with $\text{AI} > 0 \text{ km s}^{-1}$ and in the redshift interval $1.90 < z < 4.36$ (so as to contain CIV). This distribution is clearly bimodal, with one peak near 500 km s^{-1} and another around 3000 km s^{-1} .

In order to confirm and quantify the bimodality, we have applied the KMM algorithm of Ashman, Bird, & Zepf (1994). This effectively compares the quality of a single Gaussian fit to a distribution to that of a double Gaussian one. The probability that the overall log AI distribution is unimodal turns out to be negligible: the KMM likelihood test ratio statistic (essentially a χ^2) is 590 for 4 degrees of freedom. This is vastly in excess of the value of about 4 one would expect for a unimodal distribution.

The decomposition suggested by KMM is shown in the top panel of Figure 1. While there is no *a priori* reason to expect the log AI distribution to be intrinsically Gaussian (or double Gaussian), the two normal components provide quite a reasonable description of the distribution. More specifically, KMM suggests that the low-AI group contributes 49.9% of the total $\text{AI} > 0 \text{ km s}^{-1}$ population and is centered on $\text{AI} \simeq 500 \text{ km s}^{-1}$ with $\sigma \simeq 0.2$ dex; the high-AI group contributes 50.1% and is centered on $\text{AI} \simeq 3000 \text{ km s}^{-1}$ with $\sigma \simeq 0.3$ dex.

In the middle panel of Figure 1, we also show the AI distributions of all objects with $\text{BI} > 0 \text{ km s}^{-1}$ (red histogram) and of all quasars with $\text{BI} = 0 \text{ km s}^{-1}$ but $\text{AI} > 0 \text{ km s}^{-1}$ (blue histogram). This shows that the two modes exhibited by the AI-selected quasar population correspond fairly closely to “classic” BALQSOs (high-AI mode; $\text{BI} > 0 \text{ km s}^{-1}$), on the one hand, and newly added objects (low-AI mode; $\text{BI} = 0 \text{ km s}^{-1}$), on the other. The BI metric classifies 41.2% of the $\text{AI} > 0 \text{ km s}^{-1}$ objects as BALQSOs. In general, the match of the KMM-suggested groups to the $\text{BI} = 0 \text{ km s}^{-1}$ and $\text{BI} > 0 \text{ km s}^{-1}$ groups is good, except near the overlap region. The KMM decomposition suggests that the BI-selected sample may be seriously incomplete in this regime.

It should be acknowledged at this point that “bimodality” turns out to be a surprisingly slippery concept on closer examination. In particular, the number of modes in a distribution over a certain variable is not always invariant under simple transformations of that variable. This explains why the bimodality in the log AI distribution was not noticed by Trump et al. (2006), who only inspected the (linear) AI distribution. As it turns out, that distribution is, in fact, unimodal.

So does the bimodal log AI distribution actually provide evidence for two distinct QSO sub-populations? It does, because not every unimodal distribution can be transformed into a bimodal one via a logarithmic transformation. Thus while the concept of bimodality should perhaps be replaced by that of “bimodalizability” (Wyszomirski 1992), it remains true that distinct sub-populations are the most obvious way of producing such bimodalizable distributions. Indeed, in our case, the evidence for two distinct QSO sub-populations can be seen even in the linear AI distribution. In Figure 2,

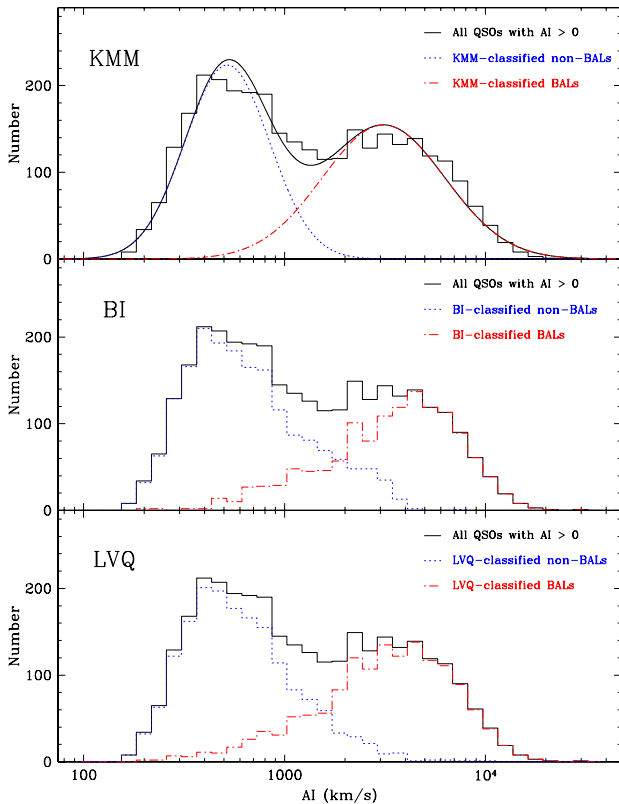


Figure 1. The log AI distribution of objects with $\text{AI} > 0 \text{ km s}^{-1}$ (black histograms in all panels). Note the obvious bimodality of this distribution. *Top panel:* The decomposition suggested by the KMM algorithm. *Middle panel:* The decomposition resulting if the classic balnicity index (BI) is used to classify BALQSOs. *Bottom panel:* The decomposition resulting if a hybrid method involving learning vector quantization (LVQ) is used to classify BALQSOs (see Section 5).

we compare the AI and log AI distribution directly. Even though the AI distribution is unimodal, it is obvious that the characteristic scale on which the distribution drops off changes abruptly at around $\text{AI} \simeq 1700 \text{ km s}^{-1}$, which coincides with the dip between the two modes of the log AI distribution. We thus believe that the evidence for two distinct sub-populations in the overall distribution is robust.²

Inspection of the linear AI distribution suggests that, beyond its mode at around $\text{AI} \simeq 400 \text{ km s}^{-1}$ (which corresponds to the low-AI mode of the log AI distribution), the drop-off in each of the two distinct regimes is roughly exponential. As shown in Figure 2, we have therefore fit a double exponential model to this distribution for $\text{AI} > 500 \text{ km s}^{-1}$. Note that, as expected, this unimodal two-

population model for the AI-distribution produces a bimodal distribution in log AI (Figure 2, top right panel). Since this double-exponential model imposes no low-AI cut-off at all on the sub-population that dominates at high-AIs, it allows us to set a useful upper limit on the size of this population (see Sections 5.3 and 6.4).

4 BEYOND STATISTICS: REPRESENTATIVE SPECTRA ACROSS THE AI/BI PARAMETER SPACE

It is important to relate the statistical results of the previous section to specific spectral properties of individual QSOs. What type of objects do we select when we apply AI and/or BI metrics, and what type of spectra correspond to different combinations of AI and BI?

The top row in Figure 3 shows the C IV line profiles of four QSOs belonging to low-AI mode in Figure 1 ($\text{AI} \sim 500$; $\text{BI} = 0 \text{ km s}^{-1}$). To our eyes, none of these QSOs appear to be genuine BALs.³ We have visually inspected the majority of similar objects and find that the same is true for most of them. Objects with $\text{AI} > 0 \text{ km s}^{-1}$ and $\text{BI} = 0 \text{ km s}^{-1}$ comprise about half of the population with $\text{AI} > 0 \text{ km s}^{-1}$, so this population is certainly not representative of “classic” BALQSOs.

The second row from the top in Figure 3 shows objects selected from the high-AI mode in Figure 1 ($\text{AI} \simeq 3000$; $\text{BI} > 0 \text{ km s}^{-1}$). As expected, all exhibit the strong and broad absorption features that are characteristic of “classic” BALQSOs.

The third row from the top in Figure 3 shows a selection of $\text{BI} = 0 \text{ km s}^{-1}$ objects from the overlap region in Figure 1 ($\text{AI} \simeq 1000 - 3000 \text{ km s}^{-1}$). It is immediately clear that these intermediate-width absorption line objects can indeed be difficult to classify with confidence. However, we have also included in this row two objects (SDSS J1730 and SDSS J1042) that appear to be genuine BALQSOs that have been missed by the BI.

The bottom row in Figure 3 shows more objects from the overlap region in Figure 1 ($\text{AI} \simeq 1000 - 3000 \text{ km s}^{-1}$), but now with $\text{BI} > 0 \text{ km s}^{-1}$. While these are clearly harder to classify than those in the high-AI mode, we think the BI has done a good job of assigning these objects to the BALQSO class.

In our view, the results of the previous and present sections imply that, although not perfect, the BI is a better metric for BALQSO identification than the AI. The majority of objects with positive BI are clearly genuine BALQSOs, but the same cannot be said with any confidence of objects classified solely on the basis of positive AI. The AI is certainly very good at finding absorbing systems, including essentially *all* BALQSOs. However, the spectroscopic properties of objects with $\text{BI} = 0 \text{ km s}^{-1}$ but $\text{AI} > 0 \text{ km s}^{-1}$, as well as the bimodality in the $\text{AI} > 0 \text{ km s}^{-1}$ population, suggest that purely AI-selected BALQSO samples will be strongly contaminated by objects with properties that

² Since our paper was accepted, Nestor, Hamann & Rodriguez Hidalgo (2008) have also found an excess of strong absorbers in a study focused mainly on relatively narrow C sc iv absorption line systems (see their Figure 9). We suspect this excess may be directly associated with the high-AI, BALQSO mode of the log AI distribution.

³ It should be acknowledged, however, that our organic neural networks have also been trained primarily on *BI-selected* BALQSOs.

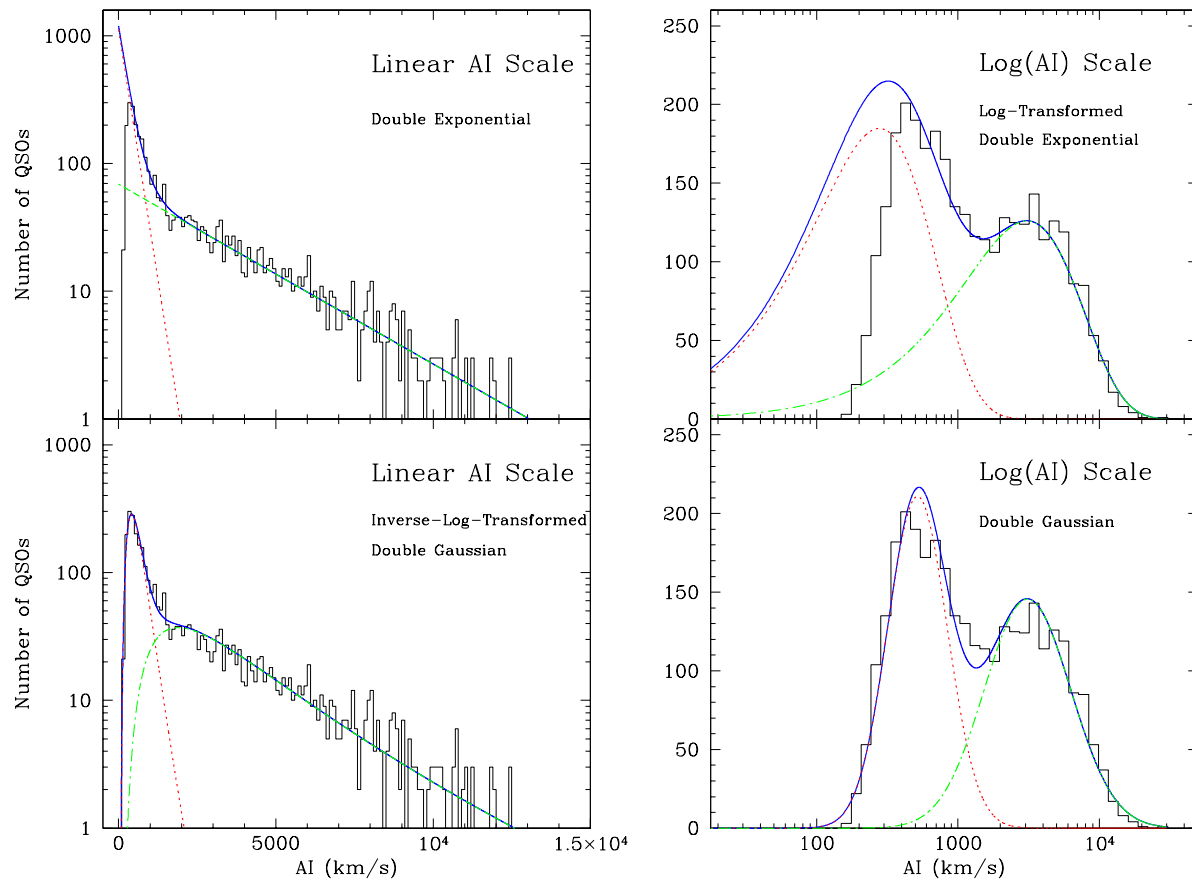


Figure 2. Comparison of the AI distribution (left panels) and the log AI distribution (right panels) of objects with $\text{AI} > 0 \text{ km s}^{-1}$. Note that only the log AI distribution is bimodal, but that the AI distribution exhibits two distinct characteristic scale lengths in the low-AI and high-AI regimes. Thus both types of distribution provide evidence of two distinct sub-populations, each of which dominates in one of these regimes. In the top panels, we also show a maximum likelihood fit to the AI-distribution above $\text{AI} = 500 \text{ km s}^{-1}$ with a double exponential model. In the bottom panels, we again show the KMM-decomposition from Figure 1, which corresponds to a double Gaussian in log AI (and a double log-normal distribution in AI).

are clearly distinct from those of “classic” BALQSOs (c.f. Ganguly et al. 2007).

The fact remains, however, that BI-selected BALQSO samples may themselves be seriously incomplete. In Section 3, we showed that the BI criterion selects 41.2% of QSOs with $\text{AI} > 0 \text{ km s}^{-1}$ as BALQSOs, whereas the KMM decomposition of the log AI distribution implies a significantly higher percentage of 50.1%. Similarly, we have now found specific examples of QSOs with $\text{BI} = 0 \text{ km s}^{-1}$ that, visually, would seem to be excellent BALQSO candidates (e.g. SDSS J1730 and SDSS J1042 in Figure 3). None of this should come as a surprise. As discussed in Section 2, there is simply no physical reason to expect that all genuine BALQSOs should have C IV absorption troughs that extend for at least 2000 km s^{-1} beyond the arbitrary 3000 km s^{-1} starting point adopted in the definition of the BI.

We conclude that BALQSO fractions derived from AI-selected samples are strong overestimates, whereas those derived from BI-selected samples are at least mild underestimates. In the following section, we will use two new methods to determine observed BALQSO fractions that are more robust than AI-based estimates and more complete than BI-based ones.

5 THE OBSERVED BALQSO FRACTION IN SDSS DR3

The fundamental problem with simple metrics such as the AI and the BI is their rigidity. For example, the BI will firmly reject an object with an absorption trough whose width is marginally less than 2000 km s^{-1} , even if this trough looks virtually indistinguishable from many objects that the BI *does* classify as BALQSOs. One way to avoid this incompleteness is to relax the classification criteria, but this incurs the danger of producing many false positives. This is what appears to have happened in the switch from the BI to the AI.

In order to overcome these problems, we have used two new approaches to estimate the observed BALQSO fraction in SDSS DR3. The first approach is based directly on the KMM-decomposition of the AI distribution in Figure 1, whereas the second approach is a hybrid method that employs a BI-trained neural network algorithm – learning-vector quantization (LVQ) – to flag potentially mis-classified objects for visual inspection. We also use a third approach – a decomposition based on the double-exponential model for the AI distribution described in Section 3 – to estimate an upper limit on the observed BALQSO fraction. The feature

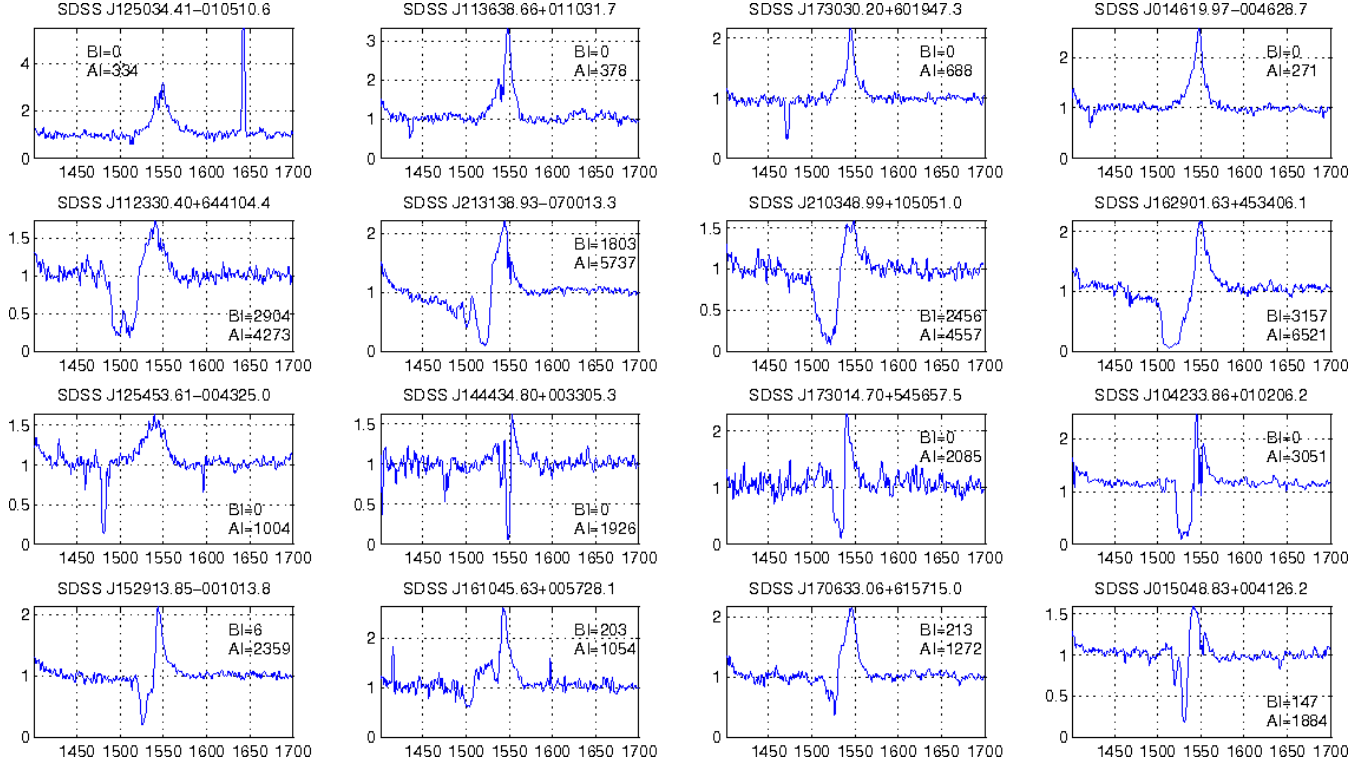


Figure 3. Representative spectra for various parts of the AI/BI parameter space. Each row corresponds to objects from a distinct region of this parameter space. *Top row:* Objects belonging to the low-AI mode in Figure 1 ($BI = 0 \text{ km s}^{-1}$; $AI \simeq 500$). *Second Row (From Top):* Objects belonging to the high-AI mode in Figure 1 ($BI > 0 \text{ km s}^{-1}$; $AI \simeq 5000$). *Third Row (From Top):* $BI = 0 \text{ km s}^{-1}$ objects belonging to the overlap region in Figure 1 ($AI \simeq 2000 \text{ km s}^{-1}$). *Bottom Row:* $BI > 0 \text{ km s}^{-1}$ objects belonging to the overlap region in Figure 1 ($AI \simeq 2000 \text{ km s}^{-1}$).

common to all three approaches is that they are fundamentally more flexible than the AI or BI metrics.

5.1 KMM-based decomposition

The KMM-based approach is straightforward. As discussed in Section 3 and shown in Figure 1 (top panel), the log AI distribution of QSOs with $AI > 0 \text{ km s}^{-1}$ can be decomposed fairly cleanly into two Gaussian components. This decomposition can be used immediately to assign a probability to each object of belonging to one or the other group. The KMM algorithm we have used provides these probabilities automatically for each object. A raw, observed BALQSO fraction can therefore be estimated from this decomposition as

$$f_{BALQSO} = \frac{1}{N_{QSO}} \sum_{i=1}^{N_{QSO}} P_{i,BALQSO} \quad (3)$$

where $P_{i,BALQSO}$ is the KMM-assigned probability that quasar i is a BALQSO (i.e. that it belongs to the high-AI mode of the distribution).

The main weakness of this method is that it assumes the KMM-decomposition to be correct. This is almost certainly not true in detail. Just as there is no reason to think that every BALQSO trough is at least 2000 km s^{-1} wide, there is no *a priori* reason to assume that the log AI distribution of BALQSOs is exactly Gaussian. However, Figure 1 suggests that a Gaussian distribution may be quite a good

approximation to the true log AI distribution of BALQSOs. The great strength of the decomposition approach is that it provides a very complete statistical census of BALQSOs (subject to its underlying assumption).

Applying this method to the full DR3 QSO sample in the redshift range $1.90 < z < 4.36$ yields an observed BALQSO fraction of $13.7\% \pm 0.3\%$ (where the error only accounts for Poisson statistics). Note that this observed global fraction is still subject to selection biases. These are dealt with in Section 6.

5.2 A hybrid method using learning vector quantization

In our second approach, we use a hybrid method to classify BALQSOs. Starting with a BI-based classification, we use a machine learning algorithm called Learning Vector Quantization (LVQ) to identify objects that might have been misclassified by the BI. All such objects are then inspected and classified visually. We will refer to this hybrid method as “LVQ-based” throughout this paper. However, it should be kept in mind that we do not use LVQ as a stand-alone BALQSO classifier, but as part of a process involving the BI, LVQ and visual inspection.

LVQ was originally devised by Kohonen (2001) and uses a neural network to assign new input data to pre-defined classes. LVQ is a particularly simple supervised neural network, in which each neuron is simply tagged as belonging to a particular class. The basic idea behind LVQ is that,

through training, each neuron should come to represent a characteristic type of object within its class. New inputs can then be assigned to classes on the basis of maximum similarity to a particular neuron.

In our case, the relevant classes are BALs vs non-BALs, and the data are continuum-normalised QSO spectra between $\lambda\lambda 1400\text{--}1700\text{ \AA}$ (spanning the C IV line). Normalization is performed exactly as in North, Knigge, & Goad (2006), and the measure of similarity we use when comparing spectra and neurons is the Euclidean distance between them (i.e. the mean rms residual). Note that we allow for redshift errors in all comparisons.

We use 800 QSOs as our training set, with 400 $\text{BI} > 0\text{ km s}^{-1}$ objects initially representing the BALQSOs and 400 $\text{BI} = 0\text{ km s}^{-1}$ objects initially representing the non-BAL QSOs. The network is then iteratively trained to classify the objects in the training set in line with their input classifications. Even though these input classifications are purely BI-based, the converged network already manages to classify some $\text{BI} = 0\text{ km s}^{-1}$ objects in the training set as likely BALQSOs (and some $\text{BI} > 0\text{ km s}^{-1}$ objects as likely non-BAL QSOs). This is possible because the network classifications are based on spectral similarity, not on the BI itself. In order to re-enforce this feature of the network, we inspect all of the “misclassified” objects in the training set visually and retag them if appropriate. We then carry out a full second training run, where the training set now includes $\text{BI} = 0\text{ km s}^{-1}$ objects explicitly tagged as BALQSOs (and vice versa). The converged network produced by this second training run is our final LVQ machine classifier. All 11,611 DR3 QSOs in the relevant redshift range are passed to this network, resulting in an LVQ classification for each of them.

As already noted above, we do not use LVQ as a stand-alone BALQSO classifier, but as a tool to flag borderline cases where the LVQ and BI classifications disagree. All such cases are then inspected and classified visually, and the visual classification is adopted as final. In practice LVQ classified 524 $\text{BI} = 0\text{ km s}^{-1}$ objects as BALQSOs, of which 334 were also classified as BALQSOs visually. Thus LVQ was quite good at identifying $\text{BI} = 0\text{ km s}^{-1}$ BALQSOs. However, LVQ also classified 383 objects with $\text{BI} > 0\text{ km s}^{-1}$ as non-BAL QSOs, and only 95 of these were also classified as non-BAL QSOs visually. This underlines the importance of the visual inspection step and justifies our reluctance to use LVQ as a stand-alone BALQSO classifier. As explained above, whenever we refer to “LVQ-based” quantities below, we will always mean quantities calculated on the basis of the full hybrid method, which uses the BI, LVQ and visual inspection.

Our LVQ-based method classifies 1,557 of the 11,611 QSOs in our DR3 parent sample as BALQSOs.⁴ The LVQ-based decomposition of $\text{AI} > 0$ objects into BALQSOs and non-BAL QSOs is shown in the bottom panel of Figure 1. The LVQ-based observed BALQSO fraction is $13.4\% \pm 0.3\%$, which is consistent with the KMM-based estimate.

5.3 Double exponential decomposition

Our third and final approach is based on the double exponential model for the (linear) AI distribution described in Section 3 and shown in Figure 2. If we associate the exponential that dominates at high-AIs with BALQSOs, we can use this model to estimate BALQSO fractions in the same way as for the KMM-based decomposition. It is worth emphasizing that this model assumes that there is no low-AI cut-off at all in the true BALQSO population – even QSOs with no absorption at all can be “BALQSOs” in this case. The observed turn-over in the AI-distribution below $\text{AI} \simeq 400\text{ km s}^{-1}$ must then be due to incompleteness. This is not entirely unreasonable, since the definition of the AI imposes a lower limit of 100 km s^{-1} and only counts absorption troughs that dip below true continuum.

While this is quite an extreme model in our opinion, it is impossible to rule out with the present data. We have therefore also estimated an “observed” BALQSO fraction on the basis of this double-exponential decomposition. This effectively provides an upper limit on the BALQSO fraction. Based on the model shown in Figure 2, we find that the exponential dominating at high-AI values corresponds to an observed BALQSO fraction of $18.3\% \pm 0.4\%$ (where the errors are again purely based on Poisson statistics). Note that this estimate includes QSOs with estimated $\text{AI} = 0\text{ km s}^{-1}$ that are not part of Trump et al.’s (2006) AI-based BALQSO catalogue. If we (somewhat arbitrarily) exclude such objects, the observed BALQSO fraction is $17.2\% \pm 0.4\%$. As explained above, we consider these estimates to be upper limits on the observed BALQSO fraction. It is therefore worth noting that even the estimate which includes $\text{AI} = 0\text{ km s}^{-1}$ objects lies substantially below the 26% BALQSO fraction suggested by Trump et al. (2006) based on the number of QSOs with $\text{AI} > 0\text{ km s}^{-1}$.

6 THE INTRINSIC BALQSO FRACTION

The observed BALQSO fractions we have derived in the previous section do not provide a fair measure of the *intrinsic* incidence of BALs within the QSO population. This is because the SDSS QSO sample suffers from a variety of selection effects that affect BALQSOs differently from non-BAL QSOs. The impact of the resulting biases can be seen in Figure 4, which shows that the observed BALQSO fractions depend strongly on redshift. As we shall see, this redshift dependence is mainly due to selection effects (c.f. Reichard et al. 2003).

In the following subsections, we first construct a more homogeneously selected QSO sample and then correct the observed BALQSO fraction derived from it for colour-, magnitude- and redshift-dependent biases. Finally, we put all of these results together to produce an unbiased estimate of the intrinsic BALQSO fraction.

6.1 A homogenous QSO parent sample

The SDSS DR3 QSO catalogue contains objects selected via a variety of selection criteria (Schneider et al. 2005). We therefore create a more homogenous QSO sample by retaining only those objects that were (or would have been)

⁴ A catalogue providing the KMM-assigned probabilities and LVQ-based classifications is available in electronic form from <http://www.astro.soton.ac.uk/~simo>.

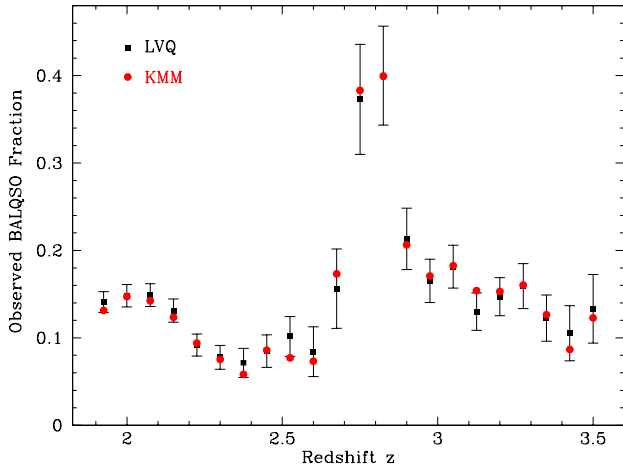


Figure 4. The redshift distribution of the *observed* BALQSO fraction. Red points correspond to the fractions determined with the KMM-based approach (see Section 5.1); black points correspond to the fractions obtained from the LVQ-based approach (see Section 5.2). No correction for selection effects has been applied to these fractions. Error bars on the KMM fractions have been suppressed for clarity, but are always similar to the LVQ ones.

selected for spectroscopic follow-up by the final QSO targeting algorithm (as described by Richards et al. 2002). This leaves us with 7,487 QSOs (out of 11,611) in our redshift range. The observed BALQSO fractions in this homogenous sample are $14.0\% \pm 0.4\%$ (KMM-based) or $14.1\% \pm 0.4\%$ (LVQ-based), but still exhibit a strong redshift dependence due to selection effects.

The SDSS QSO selection algorithm actually consists of two parallel strands, one aimed at creating a “main” QSO sample, the other aimed specifically at finding high redshift QSOs.⁵ The two strands use different limiting i' -magnitudes and colour selection criteria, which must be taken into account when dealing with the resulting selection biases. Of the 7,487 objects in our homogenous sample, 5134 would have been selected by the main sample selection criteria and 4145 by the high-redshift QSO selection criteria (1792 QSOs satisfied both sets of criteria).

6.2 Limiting-magnitude bias

There are two reasons why a magnitude cut may affect BALQSOs differently from non-BAL QSOs. First, BAL troughs may be redshifted into the bandpass where the magnitude cut is applied, causing BALQSOs to appear fainter than otherwise identical non-BAL QSOs. Second, the *continuum* spectral energy distributions (SEDs) of BALQSOs are reddened with respect to those of non-BAL QSOs. As

⁵ Strictly speaking, there is also a third strand, since objects with FIRST radio counterparts are also preferentially targeted. However, in order to correct for optical colour- and magnitude-dependent biases, we need a sample with rigorous optical selection criteria. We therefore do not include QSOs targeted solely on the basis of radio emission in our homogenous QSO sample.

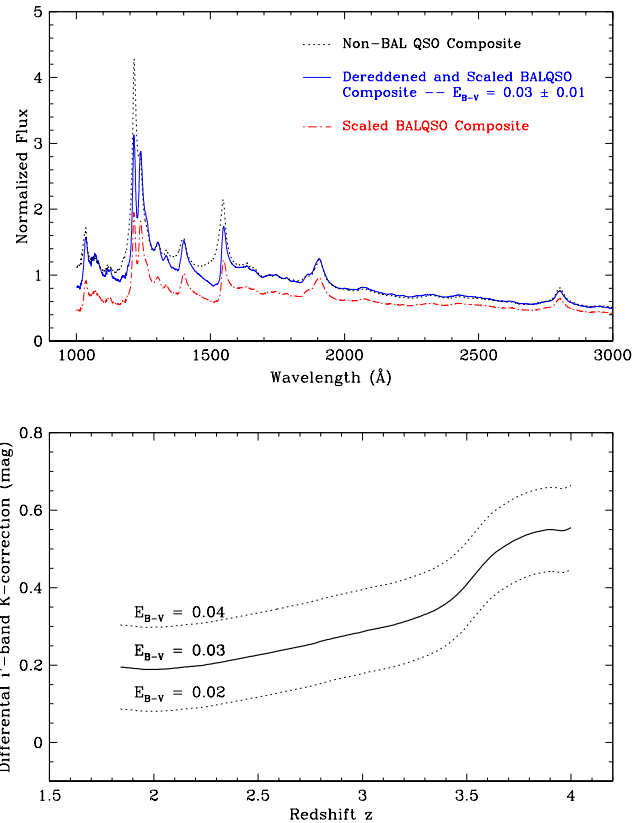


Figure 5. LVQ-based BALQSO and non-BAL QSO composites and the corresponding K-correction in the i' -band. *Top Panel:* The blue line shows the BALQSO composite after dereddening and scaling it to optimally match the non-BAL QSO composite (black line). The red line shows the original BALQSO spectrum, but scaled so that its normalization relative to the non-BAL spectrum is in line with the difference in reddening/extinction between them. *Bottom Panel:* Each line shows the redshift-dependent i' -band magnitude difference between the correctly scaled BALQSO and non-BAL QSO composites, for a particular assumed value of the differential reddening/extinction between them. The solid dark line corresponds to our preferred reddening estimate; the dotted thin lines as based on our estimate of the uncertainty on this.

already shown by Reichard et al. (2003), the form of this reddening is consistent with extinction by SMC-like dust. This again means that BALQSOs will be fainter than otherwise similar non-BAL QSOs. The consequence of these effects is that any magnitude cut will disproportionately remove BALQSOs from the sample.

In order to correct for this, we first construct BALQSO and non-BAL QSO composites and estimate the difference in reddening/extinction between them.⁶ Note that we use geometric mean composites, which ensures that the spectral index and reddening of each composite corresponds to the arithmetic mean of the spectral indices and reddening values

⁶ The composites used in this section were constructed using the LVQ-based samples; the equivalent KMM-based composites are virtually identical.

of the spectra used to construct the composite (Reichard et al. 2003). The absolute flux densities of the composites are arbitrary, however, since all individual spectra are scaled to an average value of unity in a reference wavelength interval near 1700 Å.

As shown in Figure 5 (top panel), dereddening the BALQSO composite by $E(B - V) = 0.03 \pm 0.01$ and rescaling produces a good match to the non-BAL QSO composite longward of $\simeq 1600$ Å (i.e. away from any major BAL troughs). This is consistent with the findings of Reichard et al. (2003). We therefore scale the original BALQSO composite so that its normalization relative to the non-BAL QSO composite is in line with our estimate of the difference in extinction between them (Figure 5; red line). Finally, we carry out synthetic photometry to determine the i' -band magnitude difference between BALQSOs and non-BAL QSOs as a function of redshift. This “differential K-correction” is shown in the bottom panel of Figure 5. The sharp upturn around $z \simeq 3.5$ corresponds to the first major BAL trough (C IV 1550 Å) being red-shifted into the i' -band. At the lower redshifts we will mostly be interested in below, the K-correction is only due to extinction.

We can now estimate a corrected BALQSO fraction in any redshift bin as

$$f_{BALQSO} = \frac{N_{BALQSO}}{N_{BALQSO} + N_{non-BALQSO}(i' < [i'_{lim} - \Delta i'(z)])}, \quad (4)$$

where i'_{lim} is the limiting magnitude imposed by the selection algorithm ($i'_{lim} = 19.1$ for any QSO selected only via the main sample strand; $i'_{lim} = 20.2$ for QSOs identified by the high- z colour selection). The quantity $\Delta i'(z) > 0$ is the differential K-correction. For sufficiently narrow redshift bins, this could be approximated as constant within each bin, but it is just as easy (and more precise) to calculate the K-correction independently for each non-BAL QSO according to its exact redshift. Note that N_{BALQSO} and $N_{non-BALQSO}$ become sums over probabilities when calculating f_{BALQSO} from the probabilistically-defined KMM sample (c.f. Equation 3).

Our correction for limiting-magnitude bias is similar to that applied by Hewett & Foltz (2003). It should produce reasonable results, provided that the intrinsic BALQSO and non-BAL QSO luminosity functions do not exhibit sharp breaks near the limiting absolute magnitude in any given redshift bin. One limitation of our approach is that it does not account for variations in K-correction associated with variations in BAL strength. However, BAL troughs only affect the K-correction beyond $z \gtrsim 3.5$, and in this regime we also do not have a reliable correction for colour-selection bias (see Section 6.3 and Figure 6). We therefore simply restrict our attention to lower redshifts, $z \lesssim 3.5$.

6.3 Colour-selection bias

Both the main and high- z strands of the SDSS targeting algorithm select QSO candidates on the basis of their optical photometric colours. In both strands, QSO candidates are identified as outliers from the locus defined by normal stars in the 5-dimensional SDSS colour space ($u'g'r'i'z'$). The completeness of the resulting QSO samples is a function of redshift, since even a fixed intrinsic SED produces different observed colours when placed at different redshifts.

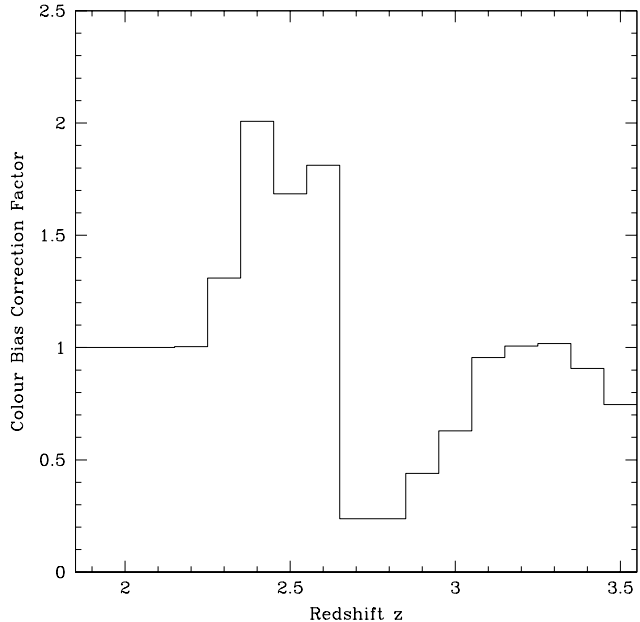


Figure 6. The redshift-dependent correction factor for colour-selection bias. This was derived by Reichard et al. (2003; see their Fig. 10), and needs to be applied to the observed BALQSO fraction.

None of this would matter for the derivation of the intrinsic BALQSO fraction if the SEDs of BALQSOs and non-BAL QSOs were identical (at least in a statistical sense). Unfortunately, they are not. First, whenever a deep BAL trough is shifted into a particular waveband, all colours involving that band are changed. Second, as discussed in Section 6.2 and shown in Figure 5, the *continuum* SEDs of BALQSOs are reddened compared to those of non-BAL QSOs. The upshot of these colour differences is that the efficiency of the SDSS QSO selection algorithm(s) is not the same for BALQSOs as for non-BAL QSOs.

Fortunately, Reichard et al. (2003) have already derived a redshift-dependent correction factor that can be applied to the observed BALQSO fraction to account for this colour-selection bias. In order to determine this correction, Reichard et al. created large sets of simulated QSO and BALQSO colours and passed both through the SDSS QSO selection algorithm. The resulting correction factor is shown as a function of redshift in Figure 6.

Three key points should be noted regarding this correction for colour-selection bias (see Reichard et al. 2003 for a full discussion). First, it is only approximate. One important limitation is that all of the simulated BALQSO colours used to derive the correction factor were based on the colour differences between a HiBALQSO composite and an average QSO composite. Thus variations in BALQSOs colours arising from the range of observed BAL strengths are not properly accounted for. It should also be kept in mind that the HiBALQSO composite used by Reichard et al. was based on a different definition of what constitutes a BALQSO than the LVQ- or KMM-based definition used here. Second, the correction factor is significantly greater than unity near $z \simeq 2.5$, but much less than unity near

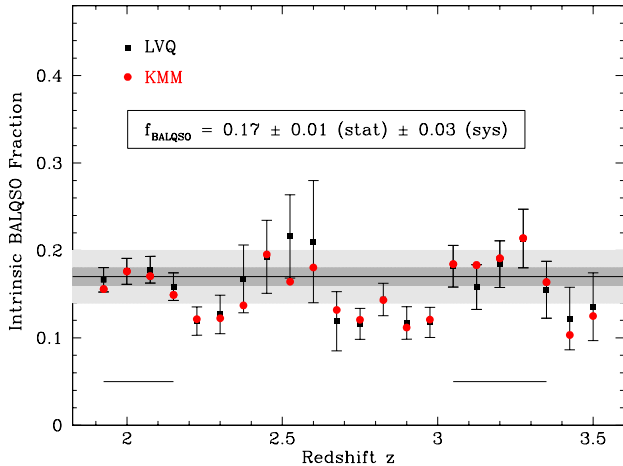


Figure 7. The redshift distribution of the *intrinsic* BALQSO fraction (after correcting for selection effects). Red points correspond to the fractions determined with the KMM-based approach (see Section 5.1); black points correspond to the fractions obtained from the LVQ-based approach (see Section 5.2). Note that the redshift dependence of the *intrinsic* fractions is markedly reduced compared to that of the *observed* fractions (c.f. Figure 4) and that the residual undulations are strongly correlated with the correction factor for colour-selection bias (Figure 6). The horizontal lines near the bottom of the plot marks redshift ranges where colour-selection bias is negligible. Our final estimate of the intrinsic intrinsic BALQSO fraction is derived from only those regions and is shown the solid horizontal line. The dark (light) shaded regions correspond to our estimate of the statistical (systematic) uncertainty on this number.

$z \simeq 2.8$. Thus the colour-selection bias causes BALQSOs to be under-represented around $z \simeq 2.5$, but over-represented around $z \simeq 2.8$ (this explains the spike at this redshift in Figure 4). Third, the correction factor is close to unity for $z \lesssim 2.2$ and $3.0 \lesssim z \lesssim 3.4$. These redshift ranges are thus optimal for estimating the intrinsic BALQSO fraction.

6.4 Putting it all together

Let us summarize all of the steps we have taken so far. First, we assigned a BALQSO or non-BAL QSO classification to every object in the redshift range $1.90 < z < 4.36$ in the SDSS DR3 QSO catalog.⁷ Next, we created a more homogenous sample by removing all objects that were not selected by the SDSS QSO targetting algorithm. We then accounted for limiting-magnitude bias by removing every non-BAL QSOs that is fainter than the effective (dereddened) magnitude limit for a BALQSO at the same redshift. Finally, we applied a redshift-dependent correction for the colour-selection bias imposed by the QSO targetting algorithm.

The final product of all these steps – and the main result of this paper – is the intrinsic BALQSO fraction plotted in Figure 7. Two key points are worth noting from this straightaway. First, the agreement between the KMM- and LVQ-based BALQSO fractions is extremely good across the

whole redshift range. This adds to our confidence that we are measuring the intrinsic abundance of a consistent class of objects. Second, the intrinsic BALQSO fractions show much less variability with redshift than the observed fractions (c.f. Figure 4), although some residual “wiggles” remain. Comparing Figures 6 and 7 immediately suggests that these wiggles are due to an imperfect correction for colour-selection bias. More specifically, the redshift dependence of the colour-correction factor is positively correlated with that of the intrinsic BALQSO fraction, so the correction derived by Reichard et al. (2003) appears to be somewhat too strong at most redshifts. We therefore do not believe that there is evidence for genuine evolution in f_{BALQSO} with redshift.

As suggested in Section 6.3, we derive our final estimate of the intrinsic BALQSO fraction from the restricted redshift ranges $1.9 < z < 2.2$ and $3.0 < z < 3.4$. These are largely free of colour-selection bias and produce consistent results. Our best estimate of the intrinsic BALQSO fraction from these regions is $f_{\text{BALQSO}} = 0.17 \pm 0.01$ (stat) ± 0.03 (sys). The statistical error here is just due to number statistics. The systematic error accounts for the uncertainty on the differential K-correction and for alternative choices in constructing the parent sample and selecting optimal redshift ranges.

We finally also estimate an upper limit on the intrinsic BALQSO fraction, based on the double exponential decomposition described in Section 5.3. The upper limit on the observed BALQSO fraction suggested by this decomposition was 18.3%, approximately 1.35 times larger than our preferred estimates of 13.7% (KMM) and 13.4% (LVQ). Since there is no evidence for a redshift dependence, we estimate an upper limit on the intrinsic fraction by applying the same factor to our best estimate of this fraction. The resulting upper limit is then $f_{\text{BALQSO}} \simeq 0.23$.

7 DISCUSSION AND CONCLUSIONS

Determining the “true” BALQSO fraction is a challenging task. A large part of the problem is the ambiguity one often encounters when attempting to classify individual absorption features as BALs or otherwise. The first goal of the present work has been to shed light on this classification problem. In this context, we have shown that when the recently introduced “absorption index” (AI) is used to classify BALQSOs, the resulting log AI distribution is clearly bimodal. Both modes contain comparable numbers of objects, but only the high-AI mode is clearly associated with genuine BALQSOs. Thus recent AI-based estimates of the BALQSO fraction – 26% (observed; Trump et al. 2006) or 43% (intrinsic; Dai, Shankar & Sivakoff 2008) – are likely to be seriously overestimated.

However, there are also good reasons to believe that the traditional “balnicity index” (BI) produces incomplete BALQSO samples. In order to make progress, we have therefore used two complementary new approaches to derive observed BALQSO fractions. One is based on a statistical decomposition of the log AI distribution, the other is a hybrid method in which a BI-trained neural network flags likely mis-identifications for visual inspection. Both approaches yield an observed BALQSO fraction around 13.5% for the SDSS DR3 QSO catalog (in the range $1.90 < z < 4.36$).

⁷ In the case of KMM, we assign a BALQSO probability.

This number should be more reliable than AI-based ones and more complete than purely BI-based ones. We also estimate an upper limit on the observed fraction of 18.3%, based on a decomposition of the AI-distribution that allows even objects without any absorption to be classified as BALQSOs.

This observed fraction is still subject to serious selection effects. We have therefore explained in detail how the observed BALQSO fraction can be corrected for colour-, magnitude- and redshift-dependent selection biases. Along the way, we confirmed that BALQSOs have redder SEDs than non-BALs, consistent with extinction by SMC-like dust at a level of $E(B - V) = 0.03 \pm 0.01$.

After applying all corrections, there is no compelling evidence for redshift evolution in the intrinsic BALQSO fraction. Our final estimate of the global intrinsic BALQSO fraction is then $f_{\text{BALQSO}} = 0.17 \pm 0.01$ (stat) ± 0.03 (sys), with an upper limit of $f_{\text{BALQSO}} \simeq 0.23$. As expected, this is similar to, but slightly higher than, the BI-based estimates from the SDSS EDR (Reichard et al. 2003). It is also similar to recent BI-based estimates (Hewett & Foltz 2003; Dai, Shankar & Sivakoff 2008) and consistent with the BALQSO fraction measured by Maddox et al. (2008) from a K-band selected QSO sample.

In closing, we would like to comment on the relationship between BALQSOs and what might be called “absorption line QSOs” (ALQSOs; this includes all objects displaying some form of absorption, such as BALs, mini-BALs, associated absorption features, narrow absorption lines...). Based primarily on the bimodality of the log AI distribution, we have argued throughout this paper that BALQSOs represent a phenomenologically distinct class amongst the ALQSOs. However, this does *not* imply that BALs and other absorption features must be produced in physically distinct line-forming regions. After all, orientation effects alone can dramatically alter the appearance of lines formed in non-spherical outflows from accretion disks (see, for example, Hamann, Korista & Morris [1993], Murray et al. [1995], or, in a different context, Knigge et al. [1995], Long & Knigge [2002]). Indeed, in the QSO unification scheme of Elvis (2000), both broad and narrow absorption lines are explicitly assumed to be formed in the same disk wind. In our view, it is likely that many, if not most, of the absorption (and perhaps also emission) line signatures seen in AGN and QSOs are formed in such accretion disk winds. We therefore agree with Ganguly & Brotherton (2008) that a comprehensive look at a wide range of outflow tracers is required in order to develop a full empirical picture of these disk winds.

The empirical distinctions between objects exhibiting different kinds of outflow tracers are important clues in this process. For example, if BALQSOs and other ALQSOs are literally “the same thing viewed from different angles”, it could be highly relevant that they occupy distinct modes of the log AI distribution. For example, in the context of orientation-based unification schemes, a restricted AI-range for BALQSOs would probably imply that the BAL-forming region of the outflow has clearly delineated physical boundaries. This would ensure that there is little room for overlap between sightlines looking into this part of the outflow (and seeing a BAL) and sightlines looking across it (and seeing only narrower absorption features). However, this conclusion cannot yet be considered robust, since different viable

decompositions of the AI distribution can produce different AI-ranges for BALQSOs.

ACKNOWLEDGEMENTS

We would like to thank Gordon Richards and Jonathan Trump for helpful responses to several questions, as well as the anonymous referee for an insightful and constructive report.

This work is supported at the University of Southampton and the University of Leicester by the Science and Technology Facilities Council (STFC).

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is <http://www.sdss.org/>.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

- Ashman K. M., Bird C. M., Zepf S. E., 1994, *AJ*, 108, 2348
- Becker, R. H. et al. 2001, *ApJS*, 135, 227
- Dai X., Shankar F., Sivakoff G. R., 2008, *ApJ*, 672, 108
- Di Matteo T., Springel V., Hernquist L., 2005, *Natur*, 433, 604
- Elvis M., 2000, *ApJ*, 545, 63
- Foltz C. B., Chaffee F. H., Hewett P. C., Weymann R. J., Morris S. L., 1990, *BAAS*, 22, 806
- Ganguly, R. et al. 2007, *ApJ*, 665, 990
- Ganguly, R. & Brotherton, M. S. 2008, *ApJ*, 672, 102
- Hall P. B., et al., 2002, *ApJS*, 141, 267
- Hamann F., Korista K. T., Morris S. L., 1993, *ApJ*, 415, 541
- Hewett P. C., Foltz C. B., 2003, *AJ*, 125, 1784
- King A., 2003, *ApJ*, 596, L27
- Knigge C., Woods J. A., Drew J. E., 1995, *MNRAS*, 273, 225
- Kohonen T., 2001, Self-organizing maps. Self-organizing maps. 3rd ed. Berlin: Springer, 2001, xx, 501 p. Springer series in information sciences, ISBN 3540679219
- Long K. S., Knigge C., 2002, *ApJ*, 579, 725

- Maddox, N., Hewett, P. C., Warren, S. J., Croom, S. M.
2008, MNRAS, in press (arXiv:0802.3650)
- Murray N., Chiang J., Grossman S. A., Voit G. M., 1995,
ApJ, 451, 498
- Nestor, D., Hamann, F. & Rodriguez Hidalgo, P. 2008, MN-
RAS, in press (arXiv:0803:0326)
- North M., Knigge C., Goad M., 2006, MNRAS, 365, 1057
- Reichard T. A., et al., 2003, AJ, 126, 2594
- Richards G. T., et al., 2002, AJ, 123, 2945
- Scannapieco E., Silk J., Bouwens R., 2005, ApJ, 635, L13
- Schneider D. P., et al., 2005, AJ, 130, 367
- Shankar F., Dai, X., Sivakoff G. R., 2008, ApJ, submitted
(arXiv:0801.4379)
- Silk J., Rees M. J., 1998, A&A, 331, L1
- Stocke J. T., Morris S. L., Weymann R. J., Foltz C. B.,
1992, ApJ, 396, 487
- Tolea A., Krolik J. H., Tsvetanov Z., 2002, ApJ, 578, L31
- Trump J. R., et al., 2006, ApJS, 165, 1
- Weymann R. J., Morris S. L., Foltz C. B., Hewett P. C.,
1991, ApJ, 373, 23
- Wyszomirski T., 1992, J. Theor. Biol., 158, 109